Stochatic gradients
Nonsmoothness, Clarke differentials, and positive homogeneity

April 1, 2021

Preliminary Definition

- Population risk $R(w) = \mathbb{E}\ell(Yf(X; w))$
- Empirical risk $\widehat{R}(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i f(x_i; w))$

Gradient descent

Let's generalize gradient descent, $w_{i+1} := w_i - \eta g_i$.

Lemma 12.1

Suppose R convex; set $G := \max_i \|g_i\|_2$, and $\eta := \frac{c}{\sqrt{t}}$. For any z,

$$R\Big(\frac{1}{t}\sum_{i<t} w_i\Big) \leq \frac{1}{t}\sum_{i<t} R(w_i) \leq R(z) + \frac{\|w_0 - z\|^2}{2c\sqrt{t}} + \frac{cG^2}{2\sqrt{t}} + \frac{1}{t}\sum_{i<t}\epsilon_t.$$

Remark

- Suppose $\|\nabla R(w_i)\| \leq G$ and set $D := max_i \|w_i - z\|$, then by Cauchy-Schwarz
$$\frac{1}{t}\sum_{i<t}\epsilon_i \leq \frac{1}{t}\sum_{i<t}\langle g_i, \nabla R(w_i), w_i - z\rangle \leq 2GD,$$

  which does not go to 0 with t.

Gradient descent

Proof

$$
\begin{aligned}
&\|w_{i+1} - z\|^2 \\
=&\|w_i - \eta g_i - z\|^2 \\
=&\|w_i - z\|^2 + 2\eta\langle g_i - \nabla R(w_i) + \nabla R(w_i), w_i - z\rangle + \eta^2\|g_i\|^2 \\
\leq&\|w_i - z\|^2 + 2\eta(R(z) - R(w_i) + \underbrace{\langle g_i - \nabla R(w_i), w_i - z\rangle}_{\epsilon_i}) + \eta^2\|g_i\|^2,
\end{aligned}
$$

which after rearrangement gives

$$2\eta R(w_i) \leq 2\eta R(z) + \|w_i - z\|^2 - \|w_{i+1} - z\|^2 + 2\eta\epsilon_i + \eta^2\|g_i\|^2,$$

and applying $\frac{1}{2\eta t}\sum_{i<t}$ to both sides gives

$$\frac{1}{t}R(w_i) \leq R(z) + \frac{\|w_0 - z\|^2 - \|w_t - z\|^2}{2\eta t} + \frac{1}{t}\sum_{i<t}(\epsilon_i + \frac{\eta}{2}\|g_i\|^2).$$

Stochastic gradient

Let us define the standard stochastic gradient oracle:

$$\mathbb{E}[g_i | w_{\le i}] = \nabla R(w_i),$$

where $w_{\le i}$ signifies all randomness in $(w_1, \ldots, w_i)$.

Remark

Sample (x, y), and set $g_i := \ell'(yf(x; w_i))y\nabla_w f(x; w_i)$; conditioned on $w_{\le i}$, the only randomness is in (x, y), and the conditional expectation is a gradient over the distribution!

Azuma-Hoeffding theorem

Suppose $(Z_i)_{i=1}^n$ is a martingale difference sequence $(\mathbb{E}(Z_i | Z_{<i}) = 0)$ and $\mathbb{E}|Z_i| \le R$. Then with probability at least $1 - \delta$,

$$\sum_i Z_i \le R\sqrt{2t \ln(1/\delta)}.$$

Stochastic gradients

Lemma 12.2

Suppose R convex; set $G := \max_i \|g_i\|_2$, and $\eta := \frac{1}{\sqrt{t}}$, $D \geq \max_i \|w_i - z\|$, and suppose g_i is a stochastic gradient at time i. With probability at least $1 - \delta$,

$$R(\frac{1}{t}\sum_{i<t} w_i) \leq \frac{1}{t}\sum_{i<t} R(w_i)$$
$$\leq R(z) + \frac{D^2}{2\sqrt{t}} + \frac{G^2}{2\sqrt{t}} + \frac{2DG\sqrt{2\ln(1/\delta)}}{\sqrt{t}}.$$

We use the above inequality to handle $\sum_{i<t} \epsilon_i$.

Stochastic gradients

Proof

Firstly, we must show the desired expectations are zero. To start,

$$\begin{aligned}\mathbb{E}[\epsilon_i | w_{\leq i}] &= \mathbb{E}[\langle g_i - \nabla R(w_i), z - w_i \rangle | w_{\leq i}] \\ &= \langle \mathbb{E}[g_i - \nabla R(w_i) | w_{\leq i}], z - w_i \rangle \\ &= \langle 0, z - w_i \rangle \\ &= 0.\end{aligned}$$

Next, by Cauchy-Schwarz and the triangle inequality,

$$\mathbb{E}|\epsilon_i| = \mathbb{E}|\langle g_i - \nabla \widehat{R}(w_i), w_i - z \rangle| \leq \mathbb{E}(\|g_i\| + \|\nabla \widehat{R}(w_i)\|)\|w_i - z\| \leq 2GD.$$

Consequently, by Azuma-Hoeffding, with probability at least $1 - \delta$,

$$\sum_i \epsilon_i \leq 2GD\sqrt{2t \ln(1/\delta)}.$$

Subgradients

Smoothness and differentials do not in general hold for us (ReLU, max-pooling, hinge loss, etc.).
One relaxation of the gradient is the **subdifferential** set ∂_s (whose elements are called **subgradients**):

$$\partial_s \widehat{R}(w) := \{s \in \mathbb{R}^p : \forall w', \widehat{R}(w') \geq \widehat{R}(w) + s^\top (w' - w)\}.$$

Typically, we lack convexity, and the subdifferential set is empty.
Our main formalism is the **Clarke differential**:

$$\partial \widehat{R}(w) := conv(\{s \in \mathbb{R}^p : \exists w_i \to w, \nabla \widehat{R} w_i \to s\}).$$

f s **locally Lipschitz** when for every point x, there exists a neighborhood $S \supseteq \{x\}$ such that f is Lipschitz when restricted to S.

Positive homogeneity

Definition

g is **positive homogeneous** of degree L when $g(\alpha x) = \alpha^L g(x)$ for $\alpha \geq 0$. (We will only consider continuous g, so $\alpha > 0$ suffices.)

Example

Layers of ReLU network are 1-homogeneous in the parameters for that layer:

$$\begin{aligned}
& f(x; (W_1, \ldots, \alpha W_i, \ldots, W_L)) \\
=& W_L \sigma(W_{L-1}\sigma(\ldots \alpha \sigma W_i \sigma(\ldots W_1 x \ldots)\ldots)) \\
=& \alpha W_L \sigma(W_{L-1}\sigma(\ldots \sigma W_i \sigma(\ldots W_1 x \ldots)\ldots)) \\
=& \alpha f(x; w).
\end{aligned}$$

The entire network is L-homogeneous in the full set of parameters:

$$\begin{aligned}
f(x; \alpha w) &= f(x; (\alpha W_1, \ldots, \alpha W_i, \ldots, \alpha W_L)) \\
&= \alpha W_L \sigma(\alpha W_{L-1}\sigma(\ldots \sigma(\alpha W_1 x)\ldots)) \\
&= \alpha^L W_L \sigma(W_{L-1}\sigma(\ldots \sigma(W_1 x)\ldots)) \\
&= \alpha^L f(x; w).
\end{aligned}$$

Positive homogeneity and the Clarke differential

Let A_i be a diagonal matrix with activations of the output after layer i on the diagonal:

$$A_i = diag(\sigma'(W_i\sigma(\ldots\sigma(W_i x)\ldots))),$$

and so $\sigma(r) = r\sigma'(r)$ implies that layer i outputs

$$x \to A_i W_i \sigma(\ldots\sigma(W_1 x)\ldots) = A_i W_i A_{i-1} W_{i-1} \ldots A_1 W_1 x.$$

The gradient with respect to layer i is

$$\frac{\mathrm{d}}{\mathrm{d}W_i} f(x; w) = (W_L A_{L-1} \ldots W_{i+1} A_i)^\top (A_{i-1} W_{i-1} \ldots W_1 x)^\top.$$

Additionally

$$\begin{aligned}
\langle W_i, \frac{\mathrm{d}}{\mathrm{d}W_i} f(x; w)\rangle &= \langle W_i, (W_L A_{L-1} \ldots W_{i+1} A_i)^\top (A_{i-1} W_{i-1} \ldots W_1 x)^\top \rangle \\
&= tr(W_i^\top (W_L A_{L-1} \ldots W_{i+1} A_i)^\top (A_{i-1} W_{i-1} \ldots W_1 x)^\top) \\
&= tr(W_L A_{L-1} \ldots W_{i+1} A_i W_i A_{i-1} W_{i-1} \ldots W_1 x) \\
&= f(x; w).
\end{aligned}$$

Positive homogeneity and the Clarke differential

Lemma 14.2

Suppose $f\colon \mathbb{R}^d \to \mathbb{R}$ is locally Lipschitz and L-positively homogeneous. For any $w \in \mathbb{R}^d$ and $s \in \partial f(w)$,

$$\langle s, w\rangle = Lf(w).$$

Proof

If $w = 0$, then $s, w = 0 = Lf(w)$ for every $s \in \partial f(w)$, so consider the case $w \neq 0$. Let D denote those w where f is differentiable, and consider the case that $w \in D \setminus \{0\}$. By the definition of gradient,

$$\lim_{\delta \downarrow 0} \frac{f(w+\delta w) - f(w) - \langle \nabla f(w), \delta w\rangle}{\delta \|w\|} = 0,$$

and by using homogeneity in the form $f(w+\delta w) = (1+\delta)^L f(w)$ (for any $\delta > 0$), then

Positive homogeneity and the Clarke differential

Proof continued

$$0 = \lim_{\delta \downarrow 0} \frac{((1+\delta)^L - 1) f(w) - \langle \nabla f(w), \delta w \rangle}{\delta}$$
$$= -\langle \nabla f(w), w \rangle + \lim_{\delta \downarrow 0} f(w)(L + O(\delta)),$$

which implies $\langle w, \nabla f(w) \rangle = L f(w)$.
Now consider $w \in \mathbb{R}^d \setminus D \setminus \{0\}$. For any sequence $(w_i)_{i \geq 1}$ in D with $\lim_i w_i = w$ for which there exists a limit $s := \lim_i \nabla f(w_i)$, then

$$\langle w, s \rangle = \lim_{i \to \infty} \langle w_i, \nabla f(w_i) \rangle = \lim_{i \to \infty} L f(w_i) = L f(w).$$

Lastly, for any element $s \in \partial f(w)$ written in the form $s = \sum_i \alpha_i s_i$ where $\alpha_i \geq 0$ satisfy $\sum_i \alpha_i = 1$ and each s_i is a limit of a sequence of gradients as above, then

$$\langle w, s \rangle = \langle w, \sum_i \alpha_i s_i \rangle = \sum_i \alpha_i \langle w, s_i \rangle = \sum_i \alpha_i L f(w) = L f(w).$$

Norm preservation

Lemma 14.3

Suppose for $\alpha > 0$, $f(x; (W_L, \dots, \alpha W_i, \dots, W_1)) = \alpha f(x; w)$. Then for every pair of layers (i, j), the gradient flow maintains

$$\frac{1}{2}\|W_i(t)\|^2 - \frac{1}{2}\|W_i(0)\|^2 = \frac{1}{2}\|W_j(t)\|^2 - \frac{1}{2}\|W_j(0)\|^2.$$

Proof.

Defining $\ell'_k(s) := y_k \ell'(y_k f(x_k; w(s)))$, and fixing a layer i,

$$\begin{aligned}\frac{1}{2}\|W_i(t)\|^2 - \frac{1}{2}\|W_i(0)\|^2 &= \int_0^t \frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\|W_i(s)\|^2 \mathrm{d}s \\ &= \int_0^t \langle W_i(s), W_i(s)\rangle \mathrm{d}s \\ &= \int_0^t \langle W_i(s), -\mathbb{E}\ell'_k(s)\frac{\mathrm{d}f(x_k; w)}{\mathrm{d}W_i(s)}\rangle \mathrm{d}s \\ &= -\int_0^t \ell'_k f(x_k; w)\mathrm{d}s.\end{aligned}$$

Norm preservation

Remark

One interesting application is to classification losses like $\exp(-z)$ and $\ln(1 + \exp(-z))$, where $\widehat{R}(w) \to 0$ implies

$$\min_k y_k f(x_k; w) \to \infty.$$

This by itself implies $\|W_j\| \to \infty$ for some j; combined with norm preservation, $\min_j \|W_j\| \to \infty$!

Smoothness inequality adapted to ReLU

Let's consider: single hidden ReLU layer, only bottom trainable:

$$f(x; w) := \frac{1}{\sqrt{m}} \sum_j a_j \sigma(\langle x, w_j \rangle), \qquad a_j \in \{+1, -1\}.$$

Let $W_s \in \mathbb{R}^{m \times d}$ denote parameters at time s, suppose $\|x\| \leq 1$.

$$\frac{\mathrm{d}f(x; W)}{\mathrm{d}W} = \begin{bmatrix} a_1 x \sigma'(w_1^\top x)/\sqrt{m} \\ \vdots \\ a_m x \sigma'(w_m^\top x)/\sqrt{m} \end{bmatrix},$$

$$\|\frac{\mathrm{d}f(x; W)}{\mathrm{d}W}\|_F^2 = \sum_j \|a_j x \sigma'(w_j^\top x)/\sqrt{m}\|_2^2 \leq \frac{1}{m} \sum_j \|x\|_2^2 \leq 1.$$

Smoothness inequality adapted to ReLU

We will use the logistic loss, whereby

$$\ell(z) = \ln(1 + \exp(-z)),$$
$$\ell'(z) = \frac{-\exp(-z)}{1 + \exp(-z)} \in (-1, 0),$$
$$\hat{R}(W) := \frac{1}{n} \sum_k \ell(y_k f(x_k; W)).$$

A key fact is $|\ell'(z)| = -\ell'(z) \leq \ell(z)$, whereby

$$\frac{\mathrm{d}\widehat{R}}{\mathrm{d}W} = \frac{1}{n} \sum_k \ell'(y_k f(x_k; W)) y_k \nabla_W f(x_k W),$$
$$\|\frac{\mathrm{d}\widehat{R}}{\mathrm{d}W}\|_F \leq \frac{1}{n} \sum_k |\ell'(y_k f(x_k; W))| \cdot \|y_k \nabla_W f(x_k W)\|_F$$
$$\leq \frac{1}{n} \sum_k |\ell'(y_k f(x_k; W))| \leq \min\{1, \widehat{R} W\}.$$

Smoothness inequality adapted to ReLU

Lemma 14.4

If $\eta \leq 1$, for any Z,

$$\|W_t - Z\|_F^2 + \eta \sum_{i<t} \widehat{R}^{(i)}(W_i) \leq \|W_0 - Z\|_F^2 + 2\eta \sum_{i<t} \sum_{i<t} \widehat{R}^{(i)}(Z),$$

where $\widehat{R}^{(i)}(W) = \frac{1}{n} \sum_k \ell(y_k < W, \nabla f(x_k; W_i) >)$.

Proof

Using the squared distance potential as usual,

$$\|W_{i+1} - Z\|_F^2 = \|W_i - Z\|_F^2 - 2\eta\langle \nabla \widehat{R}(W_i), W_i - Z\rangle + \eta^2 \|\nabla \hat{R}(W_i)\|_F^2,$$

where $\|\nabla \widehat{R}(W_i)\|_F^2 \leq \|\nabla \widehat{R}(W_i)\|_F \leq \widehat{R}(W_i) = \widehat{R}^{(i)}(W_i)$, and

Smoothness inequality adapted to ReLU

Proof continued

$$
\begin{aligned}
& n\langle \nabla \widehat{R}(W_i), Z - W_i\rangle \\
=& \sum_k y_k \ell'(y_k f(x_k; W_i)) \langle \nabla_W f(x_k; W_i), Z - W_i\rangle \\
=& \sum_k \ell'(y_k f(x_k; W_i)) (y_k \langle \nabla_W f(x_k; W_i), Z\rangle - y_k f(x_k; W_i)) \\
\leq& \sum_k (\ell(y_k \langle \nabla_W f(x_k; W_i), Z\rangle) - \ell(y_k f(x_k; W_i))) \\
=& n(\widehat{R}^{(i)}(Z) - \widehat{R}^{(i)}(W_i)).
\end{aligned}
$$

Together,

$$\|W_{i+1} - Z\|_F^2 \leq \|W_i - Z\|_F^2 + 2\eta(\widehat{R}^{(i)}(Z) - \widehat{R}^{(i)}(W_i)) + \eta \widehat{R}(W_i);$$

applying $\sum_{i<t}$ to both sides gives the bound.